Direct Effects Testing: A Two-Stage procedure to Test for Effect Size and Variable Importance for Correlated Binary Predictors and a Binary Response

M. SPERRIN T. JAKI

June 10, 2009

Abstract

In applications such as medical statistics and genetics, we encounter situations where a large number of highly correlated predictors explain a response. For example, the response may be a disease indicator and the predictors may be treatment indicators or single nucleotide polymorphisms (SNPs). Constructing a good predictive model in such cases is well studied. Less well understood is how to recover the 'true sparsity pattern', that is finding which predictors have direct effects on the response, and indicating the statistical significance of the results. Restricting attention to binary predictors and response, we study the recovery of the true sparsity pattern using a two-stage method that separates establishing the presence of effects from inferring their exact relationship with the predictors. The uncertainty in the relationship between the predictors and the recovered effects is represented by a discrete distribution giving the likelihood of the effect originating from each of a collection of predictors. Simulations and a real data

application demonstrate the method discriminates well between associations and direct effects. Comparisons with lasso based methods demonstrate favourable performance of the proposed method.

Keywords: Contingency table; Direct effect; High Dimensional; Lasso; Noncentral hypergeometric distribution; Sparsity.

1 Introduction

It is commonplace in applications of statistics to encounter situations in which a large number of predictors are available to explain a response. Consider the classical regression

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is an $n \times 1$ response vector explained by an $n \times p$ design matrix X through an unknown $p \times 1$ coefficient vector β with $n \times 1$ noise vector ε. Having a large number of predictors, p, possibly even $p > n$, should intuitively be beneficial, as we are maximising the information available to explain the response. From the perspective of producing a good predictive model, this is true, and many methods are available for this objective, such as principal component regression [Massey, 1965], partial least squares [Wold, 1975], ridge regression [Hoerl and Kennard, 1988] and more recent methods such as sparse sufficient data reduction [Li, 2007].

In this paper we are interested in recovering the so-called 'true sparsity pattern' [Wasserman and Roeder, 2009], in which we search for a subset of predictors deemed to have a 'direct

effect' on the response — that is an effect that is causally attributed to the predictor in question rather than being due to the correlation of the predictor with other important predictors. We wish to find a sparse solution to the regression given in Equation (1) and in particular carry out significance tests of variable importance. The lasso [Tibshirani, 1996] is a very popular sparse estimator, where sparsity is induced by applying an L_1 penalty to the size of the vector β. It is computationally fast thanks to the least angle regression algorithm (LARS) of Efron et al. [2004]. Other possibilities for sparse estimation include subset selection [Breiman, 1995], the Dantzig selector [Candes and Tao, 2007] and sure independence screening [Fan and Lv, 2008]. For the lasso, much work has been carried out concerning consistency in terms of sparse pattern recovery [see for example Knight and Fu, 2000, Zou, 2006, Bunea et al., 2007].

Until recently, it has not been possible to reliably ascertain significance of parameters included in a sparse model, that is to test for variable importance. Although standard errors of lasso parameters are available [Tibshirani, 1996, Osborne et al., 1998] these are difficult to interpret because of the discontinuity of the sampling distribution of the parameters. In the situation where the predictors in the model are not too highly correlated, recent methods that address this include the 'screen and clean' method [Wasserman and Roeder, 2009, Meinshausen et al., 2008], and stability selection [Meinshausen and Bühlmann, 2008]. Such methods are also appropriate when, in the highly correlated predictor case, it is satisfactory to recover predictors that are correlated with those that are truly causal. However, carrying out significance tests in the presence of multicollinearity is, according to Meinshausen [2008], p266, 'in some sense ill-posed'.

There are many situations, however, when multicollinearity can be serious, and we are interested nevertheless in recovering the true sparsity pattern, along with ascertaining the significance of our result. For example, in genomewide association studies we study a number of sites on the genome called single nucleotide polymorphisms (SNPs) which are highly correlated with each other. We would like to identify exact regions on the genome that influence the risk of disease, so that appropriate interventions can be considered. The problem of multicollinearity can be seen by considering a group $\mathcal{J}$ of highly correlated predictors, one of which has a true non-zero regression coefficient (or direct effect). Then the lasso tends to select one variable from $\mathcal{J}$, but there is no stability in which variable is selected. This is noted by Zou and Hastie [2005], who propose as a solution the 'elastic net', which modifies the lasso by adding an L_2 penalty, that promotes inclusion of all the predictors in the group $\mathcal{J}$. Whilst this improves the sensitivity of recovering the sparsity pattern, this is at the expense of inclusion of a potentially large number of noise predictors in the model, and effect sizes becoming difficult to interpret because they are 'shared' amongst the correlated predictors. Such an approach is useful, for example, in the recovery of gene networks, but not for the true sparsity recovery problem considered here. Meinshausen [2008] adopts a hierarchical approach, in which he looks for significance at the level of groups of variables, rather than the level of individual variables. This is sensible, since in the case of the group $\mathcal{J}$ of highly correlated predictors, it can be easy to identify that at least one member of the group has a direct effect, but difficult or impossible to identify which member(s) of the group have the effect. However, the method relies upon the selection of an appropriate hierarchical clustering regime, and it is apparent that the results will depend

upon the clustering method chosen.

In this paper we introduce a two-stage method that allows separation of the two inherent kinds of uncertainty: presence of an effect (sufficiently large to be deemed significant) and which predictor(s) the effect is allied to. The application of the method is to 'fine mapping' problems — those where the correlation is particularly high — and in particular may violate the standard correlation structure assumptions relied upon by other methods for consistency results [see Meinshausen and Bühlmann, 2008, for a summary of these assumptions and further references]. Consequently, our method makes no claims about consistency of variable selection. Instead, the idea is to acknowledge uncertainty about which predictor is the source of a given effect by providing probabilities that a direct effect arises from each of a collection of predictors. Currently, we restrict attention to binary predictors and response. The key element of the method is a novel recasting of the regression problem as

$$Z = EM + \epsilon, \tag{2}$$

where Z is a $p \times 1$ vector constructed to represent the marginal association of each predictor with the response, M is an unknown $p \times 1$ vector containing the direct effect of each predictor with the response, and E is a $p \times p$ effect matrix constructed to translate the direct effects into the observed associations, by considering the correlation structure of the predictors. These objects are formally defined in Section 2. We estimate M via lasso regression [Tibshirani, 1996], where Z is taken as the response and E the design matrix, to give a collection of direct effects that are coherent with the observed association structure of all the predictors with the response. We then separately consider the uncertainty of M

in terms of the size of the effect, and which predictor is linked to the effect. The main advantage of considering the regression in Equation (2) rather than Equation (1) is that, under some assumptions, distributions for the effect size, not influenced by multicollinearity, are readily available. The output of the method is then a collection of significant direct effects, each with a probability distribution expressing the uncertainty in the associated predictor across a set of predictors. We call the method direct effect testing (DET).

The method is similar in spirit to Meinshausen [2008] in that we identify significant effects but acknowledge uncertainty about the specific predictors involved, but here we are able to specify relative confidence in each predictor being the origin of a given effect. Also, whilst the method of Meinshausen [2008] can be considered a 'top down' approach, starting out with large clusters, and gradually moving down the hierarchy to smaller clusters, our method works in the opposite direction, since we test on an individual predictor level for effects, then generate a cluster that contains potential predictors for the true origin of a given effect.

In the remainder of the manuscript, we formally define the methodology in Section 2, before we investigate its behaviour on simulated data in Section 3 and real data in Section 4. We conclude with a summary and discussion in Section 5.

Table 1: Notation for a 2×2 contingency table for a binary response Y and binary predictor X_j

	Observed Counts		
	$Y = 0$	$Y = 1$	Total
$X_j = 0$	a_j	b_j	t_{0j}
$X_j = 1$	c_j	d_j	t_{1j}
Total	s	r	n

2 Method

2.1 Definitions and Notation

Suppose we are interested in a binary response Y, and its relationship to a set of p binary predictors $\mathbf{X} = (X_1, \ldots, X_p)$. Consider the situation where we have n complete observations of the form $(y_i, x_{i1}, \ldots, x_{ip}) \in \{0, 1\}^{p+1}$, $i = 1, \ldots, n$. Table 1 gives further notation that will be used. Without loss of generality we assume in the sequel that $\text{cor}(X_j, Y) \geq 0$, $j = 1, \ldots, p$, reversing the binary coding for X_j whenever this does not hold.

We will use the language of graph theory to introduce the concept of direct effects — see, for example, Pearl [2009] for an introduction. Consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Let the vertices correspond to the $p + 1$ binary variables $\{Y = X_0, X_1, \ldots, X_p\}$, and the edges correspond to dependencies between the vertices. With the understanding that $X_0 = Y$, the edge (j_1, j_2) is absent if and only if X_{j_1} and X_{j_2}

are conditionally independent given all the other variables, i.e.

$$X_{j_1} \perp\!\!\!\perp X_{j_2} | X_{-(j_1,j_2)},$$

where $X_{-(j_1,j_2)}$ means all the variables except X_{j_1} and X_{j_2}.

We are interested in dependencies between the variables $\mathbf{X}$ and the response Y. A *path* is an unbroken sequence of edges through the graph; two variables are *connected* if there exists a path between them. If X_j and Y are connected they are not independent, and hence associated.

A hypothesis test of this association is

$$\begin{aligned} H_0^j &: \quad X_j \text{ is not associated with } Y, \\ H_1^j &: \quad X_j \text{ is associated with } Y. \end{aligned}$$

Since all the variables are binary, the null hypothesis H_0^j implies that the count a_j in the contingency table (Table 1) is distributed according to a hypergeometric distribution with mean μ_{0j} and variance σ_{0j}^2, given by

$$\begin{aligned} \mu_{0j} &= \frac{st_{0j}}{n}, \\ \sigma_{0j}^2 &= \frac{rst_{0j}t_{1j}}{n^2(n-1)}. \end{aligned}$$

Therefore, such a hypothesis test can be carried out using Fisher's exact test for each $X_j, j = 1, \ldots, p$.

Two variables are *adjacent* if there exists an edge between them (i.e. they are connected by a path of length one). If X_j and the response Y are adjacent we say there is a *direct*

effect between X_j and Y. If there exists a path of length two between X_j and Y, we say there is an *indirect effect* between X_j and Y. We ignore any path of length greater than two. There may be numerous indirect effects between X_j and Y, and direct and indirect effects can co-exist.

A hypothesis test of a direct effect is

$$\begin{aligned} \tilde{H}_0^j &: X_j \text{ is not directly affecting } Y, \\ \tilde{H}_1^j &: X_j \text{ is directly affecting } Y. \end{aligned}$$

Regardless of which of the above hypotheses apply, the count a_j is distributed according to Fisher's noncentral hypergeometric distribution [McCullagh and Nelder, 1989] with, say, mean $\tilde{\mu}_{\omega j}$ and variance $\tilde{\sigma}^2_{\omega j}$, under $\tilde{H}^j_\omega$, $\omega = 0, 1$. Under $\tilde{H}^j_1$ the noncentrality of the distribution is allowed to include a potential direct effect between X_j and Y, but under $\tilde{H}^j_0$ the noncentrality accounts for indirect effects only. For convenience we will drop the subscript ω when we talk about the noncentral hypergeometric distribution in general.

The mean, $\tilde{\mu}_j$, of Fisher's noncentral hypergeometric distribution is available when the noncentrality of the distribution is known [McCullagh and Nelder, 1989]. In this application, however, the noncentrality is not known as it depends on the potential association of each predictor X_j with Y. Once the mean is known, the variance can be approximated by [Levin, 1984]:

$$\begin{aligned} \tilde{\sigma}^2_j &\approx \frac{ngh}{(n-1)(t_{0j}h + t_{1j}g)}, \\ g &= \tilde{\mu}_j(t_{0j} - \tilde{\mu}_j), \quad h = (s - \mu_j)(\mu_j + t_{1j} - s). \end{aligned} \tag{3}$$

The mean of the noncentral distribution $\tilde{\mu}_j$ can be written as the sum of the standard

hypergeometric mean μ_{0j} and some function of the noncentrality. We propose modelling the noncentrality part explicitly as a linear combination of the direct and indirect effects between X_j and Y.

2.2 Noncentrality Model

The first stage in constructing the direct effect testing model is to estimate the direct and indirect effects in a coherent framework that reflects the correlation structure of the dataset. Let $z_j = \sigma_{0j}^{-1}(a_j - \mu_{0j})$, so that z_j is the count a_j standardized to have zero mean and unit variance under H_0^j. Either a_j or z_j could be used to test an association hypothesis between X_j and Y.

Next, define the event $C_k = \{\text{Only predictor } X_k \text{ has a direct effect with } Y\}$. Under C_k, any path in the graph $\mathcal{G}$ from Y to any X_j with $j \neq k$ must pass through X_k — we say Y is *separated* from X_{-k}. As a consequnce,

$$Y \perp\!\!\!\perp X_j | X_k$$

for all $j \neq k$. Let $e_j^k = E(z_j \mid C_k, z_k)$, which is the regression function of z_j on z_k, under the condition C_k. Then clearly $e_j^j = z_j$ for each j, while for the general case, straightforward algebra (see Appendix) can be used to derive

$$e_j^k = \sigma_{0j}^{-1}\left\{ n\left(\frac{\gamma_{0,0}a_k}{a_k + b_k} + \frac{\gamma_{0,1}c_k}{c_k + d_k}\right) - \mu_{0j}\right\}, \tag{4}$$

where

$$\gamma_{\omega_1,\omega_2} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(x_{ik} = \omega_1, x_{ij} = \omega_2), \tag{5}$$

with $\mathcal{I}(\cdot)$ denoting the indicator function.

We now model z_j, $j = 1, \ldots, p$, as a linear combination of its indirect effect induced by the other predictors, its own direct effect, and the residual noise, which can be written as

$$z_j = \sum_{k=1}^{p} m_k e_j^k + \epsilon_j, \tag{6}$$

given in vector form in Equation (2). In Equation (6), each m_k denotes the direct effect between predictor X_k and Y. We expect most of these to be zero, and $m_k \neq 0$ means that predictor X_k has a direct effect on the response Y. Also, if $m_k = 0$ this corresponds to the truth of the hypothesis $\tilde{H}_0^k$. Since all $m_k,\ k = 1, \ldots, p$ are unknown we will estimate them by $\hat{m}_k,\ k = 1, \ldots, p$ via lasso regression [Tibshirani, 1996], using the least angle regression algorithm [Efron et al., 2004]. In order to choose the constraint on the lasso, note that because $E(\epsilon_j) = 0$ for each j,

$$\sum_{j=1}^{p} \text{var}(\epsilon_j) = E\left(\epsilon_j^2\right) = \sum_{j=1}^{p} \frac{\tilde{\sigma}_{1j}^2}{\sigma_{0j}^2}. \tag{7}$$

We therefore select the point on the lasso path where $\sum_{j=1}^{p} \epsilon_j^2$ is equal to its expectation (Equation 7). The noncentral variance $\tilde{\sigma}_{1j}^2$ depends upon the current noncentrality estimate, hence is recalculated for every step along the lasso path. We make no assumption about the presence or absence of direct effects at this stage — this is controlled by the estimate $\hat{M}$.

The model (6) is not homoskedastic because $\text{var}(\epsilon_j) = \sigma_j^2/\sigma_{0j}^2$, so the variances depend on the size of the noncentrality of each predictor X_j. However, scaling by the standard deviation under each H_0^j provides some stability. Furthermore, the more severe the noncentrality of X_j, the smaller its variance tends to be, so there will not be points that exert excessive leverage on the linear model due to large variances. Also, the ϵ_js are not inde-

pendent, they are partially determined through the correlation structure of the predictors. Classical regression carried out in a situation of non-independent errors leads to coefficient estimates that are still unbiased, but are unlikely to be the best linear unbiased estimator.

Ideally, we would like to carry out the hypothesis tests $(\tilde{H}_0^j, \tilde{H}_1^j)$ to establish whether or not a direct effect exists between X_j and Y, for each $j = 1, \ldots, p$. For a given j, if we knew the direct effects on the other predictors, M_{-j}, we could calculate the indirect effect between X_j and Y, and hence the noncentrality of the noncentral hypergeometric null distribution. Then, the distribution of a_j under $\tilde{H}_0^j$ would have mean $\tilde{\mu}_{0j}$ and variance $\tilde{\sigma}_{0j}^2$ where

$$\tilde{\mu}_{0j} = \mu_{0j} + \sigma_{0j} \sum_{k \neq j} m_k e_j^k.$$

This comes about by taking the central mean μ_{0j}, and estimating the null noncentrality parameter as a linear combination of all the indirect effects between X_j and Y, and then $\tilde{\sigma}_{0j}^2$ is estimated via Equation (3). Thus any remaining association can be attributed to a direct effect.

Unfortunately, we only have an estimate $\hat{M}$, and hence we cannot carry out the above hypothesis tests explicitly. We therefore resort to a two-stage procedure in which we separate the uncertainty in $\hat{M}$ into effect size uncertainty and predictor assignment uncertainty.

2.3 Stage One — Hypothesis Testing for Effect Size

Recall that for a set $\mathcal{J}$ of highly correlated predictors, one of which has a direct effect, the lasso tends to select one variable from the group, but there is no stability in which variable is selected. Therefore, the coefficient estimate $\hat{m}_j$ assigned to predictor X_j can be used to

estimate the *size* of the corresponding effect, but we must bear in mind that X_j may not be the actual predictor from which the effect originates — it may be one of its neighbours in $\mathcal{J}$. We test for significance of the size of the effect assigned by the lasso to each predictor using a Fisher's noncentral hypergeometric null distribution with the estimate $\hat{M}$ plugged in. Denoting the resulting mean by $\hat{\tilde{\mu}}_{0j}$ and the variance by $\hat{\tilde{\sigma}}^2_{0j}$,

$$\hat{\tilde{\mu}}_{0j} = \mu_{0j} + \sigma_{0j} \sum_{k \neq j} \hat{m}_k e_j^k,$$

and again the variance is estimated via Equation (3). The test statistic is then calculated as

$$T = \frac{z_j - \hat{\tilde{\mu}}_{0j}}{\hat{\tilde{\sigma}}_{0j}},$$

and this can either be tested against the relevant non-central hypergeometric distribution, or provided the margins of the contingency table are sufficiently large, an approximation to a standard normal distribution is possible. It is interesting that $\hat{m}_j = 0$ could still lead to the effect assigned by the lasso to X_j being deemed significant. The multiple testing issue arising at this point can be addressed using one's favourite method of error control — we have simply used a Bonferroni correction in this paper.

2.4 Stage Two — Uncertainty in Direct Effect Predictor Assignment

Suppose predictor X_j has a direct effect on the response Y, but is highly correlated with predictor X_k. Then by chance it may happen that $\text{cor}(X_k, Y) > \text{cor}(X_j, Y)$, and thus the lasso wrongly identifies the effect on predictor X_k [see also Zou and Hastie, 2005]. For each detected effect, we therefore identify a class of predictors from which each effect could

truly have originated. Moreover, we allocate a probability to each predictor in this class measuring the likelihood that the effect originated from that predictor. Returning to the graph theory analogy, in the first stage we have established the number of edges originating from the response Y, and roughly where each edge leads. We now acknowledge uncertainty, over a small set of vertices, for each edge.

When an effect is declared, in stage one, on a predictor X_k, we generate a class $\{X_j : j \in \mathcal{J}\}$ of predictors highly correlated with X_k (including X_k itself). Then for each $j \in \mathcal{J}$ we would like to calculate $p_{j|k} = \text{pr}(X_j \text{ true direct effect}|X_k \text{ declared direct effect})$.

To proceed we use the result that

$$\begin{aligned} p_{j|k} \quad \propto \quad & \text{odds}(X_k \text{ declared DE}|X_j \text{ true DE}, X_j \text{ or } X_k \text{ declared DE}) \\ & \times \text{pr}(X_j \text{ declared DE}|X_j \text{ true DE})\text{pr}(X_j \text{ true DE}), \end{aligned} \tag{8}$$

where DE stands for direct effect. A proof is given in the Appendix. We make three assumptions in the sequel:

1. The set $\mathcal{J}$ covers all reasonable predictors, in that $p_{j|k}$ is negligible for any $j \notin \mathcal{J}$. We discuss the choice of $\mathcal{J}$ at the end of this section.

2. Each predictor is *a-priori* equally likely to be responsible for a direct effect on Y.

3. For each $j \in \mathcal{J}$, $\text{pr}(X_j \text{ declared DE}|X_j \text{ true DE})$ is the same. In other words the sensitivity of the method does not depend on which predictor happens to possess the effect.

These assumptions allow us to calculate $p_{j|k}$ for each $j \in \mathcal{J}$ as

$$p_{j|k} = \frac{\text{odds}(X_k \text{ declared DE}|X_j \text{ true DE, } X_j \text{ or } X_k \text{ declared DE})}{\sum_{l\in\mathcal{J}} \text{odds}(X_k \text{ declared DE}|X_l \text{ true DE, } X_l \text{ or } X_k \text{ declared DE})}. \quad (9)$$

We now outline the procedure for calculating the right hand side of Equation (9). Suppose an effect has been observed in stage one between X_k and Y. Let β_k be the size of the direct effect, measured as the change in the estimated effect size if X_k were changed from $\{X_k = 0\}$ to $\{X_k = 1\}$, but all other variables X_{-k} were held constant, and let α_k be the baseline effect size under $\{X_k = 0\}$, with the other variables unchanged, so that

$$\begin{aligned} \beta_k &= \text{pr}(\text{Y}_{\{\text{X}_\text{k}=1\}} = 1) - \text{pr}(\text{Y}_{\{\text{X}_\text{k}=0\}} = 1), \\ \alpha_k &= \text{pr}(\text{Y}_{\{\text{X}_\text{k}=0\}} = 1). \end{aligned} \quad (10)$$

We estimate α_k and β_k using the association measure z_k, with the indirect effects removed,

$$\begin{aligned} \hat{\alpha}_k &= \frac{t_{0k} - \mu_k - \sigma_k(z_k - \sum_{k\neq j} m_k e_j^k)}{t_{0k}}, \\ \hat{\beta}_k &= \frac{r - t_{0k} + \mu_k + \sigma_k(z_k - \sum_{k\neq j} m_k e_j^k)}{t_{1k}} - \hat{\alpha}_k, \end{aligned} \quad (11)$$

see the Appendix for further details.

Suppose that X_j has a true direct effect on Y, but this effect has been detected, in stage one, on predictor X_k. The effective number of observations that we can use to distinguish between X_j and X_k as the origin of the effect is given by

$$N_E(j,k) = n(\gamma_{(0,1)} + \gamma_{(1,0)}),$$

i.e. when the two predictors take different values. Evidence towards X_j rather than X_k truly possessing direct effect, the 'truth' in this case, occurs when $(X_j, X_k, Y) = (0,1,0)$ or $(1,0,1)$.

Suppose this happens $ET(j,k)$ times. Evidence towards predictor k rather than predictor j having a direct effect, the incorrect conclusion, occurs when $(X_j, X_k, Y) = (0,1,1)$ or $(1,0,0)$. Suppose this happens $EF(j,k)$ times. It is clearly possible to observe $EF(j,k) > ET(j,k)$, and is particularly likely for small β_j, small n or large correlation between X_j and X_k, resulting in the aforementioned scenario, that X_k is wrongly detected as possessing the direct effect.

Using straightforward algebra (see Appendix),

$$\begin{aligned} P_{EF(j,k)} &= \text{pr}\left[(\mathrm{X_j}, \mathrm{X_k}, \mathrm{Y}) = (0,1,1) \text{ or } (1,0,0) \mid \mathrm{X_j} \neq \mathrm{X_k}\right] \\ &= \frac{\gamma_{(1,0)}\alpha_k + \gamma_{(0,1)}(1-\alpha_k-\beta_k)}{\gamma_{(1,0)} + \gamma_{(0,1)}}, \end{aligned} \tag{12}$$

with $\gamma(\omega_1, \omega_2)$ as in Equation (5). For intuition, note that if we assume $t_{0j} = t_{0k}$ this reduces to

$$\hat{P}_{EF(j,k)} = \frac{1-\beta_k}{2}.$$

It follows that

$$EF(j,k) \sim \text{Binomial}(N_E(j,k), P_{EF(j,k)}), \tag{13}$$

so we can use this to calculate, for each $j \in \mathcal{J}$,

$$\text{pr}\{EF(j,k) > ET(j,k)\} = \text{pr}(EF > N_E/2).$$

However, note the equality of events

$$\{EF(j,k) > ET(j,k)\} = \{k \text{ declared DE} \mid j \text{ true DE}, j \text{ or } k \text{ declared DE}\} \tag{14}$$

that comes about as a consequence of the behaviour of the lasso. Hence, recalling Equation (9), this gives us a mechanism to calculate $p_{j|k}$ for $j \in \mathcal{J}$.

There are various ways that $\mathcal{J}$ could be chosen. A cut-off value of ρ could be found so that, where ρ_{jk} is the correlation between X_j and X_k, pr{EF(j, k) > ET(j, k)} is small for $\hat{\rho}_{jk} < \rho$, i.e. X_j is very unlikely to be the true causal predictor associated with X_k. Alternatively, one could fix the size of $\mathcal{J}$ to, say, the ten predictors that are most highly correlated with X_k; or in the spirit of Meinshausen [2008], one could consider using clustering algorithms to select $\mathcal{J}$. In the subsequent work, we adopt the first approach, and choose ρ such that $\text{pr}\{\text{EF(j, k)} \geq \text{ET(j, k)} \mid \hat{\rho}_{\text{jk}} < \rho\} \leq 0{\cdot}01$. Practically, provided conservative bounds are selected when choosing $\mathcal{J}$ the choice of the set is not important. Indeed, one could simply allow $\mathcal{J}$ to contain all the predictors, in this case those predictors that are not highly correlated with X_k would turn out to have a negligible probability of containing the true direct effect.

3 Simulated Data

We will now evaluate direct effect testing on the 'ge03d2' dataset taken from the 'GenABEL' package [Aulchenko and Struchalin, 2008] in R [R Development Core Team, 2008]. This dataset contains $n = 897$ subjects, with $p = 7480$ SNPs measured on each subject. We restrict our attention to dominant effects of the SNPs so that, in the usual coding of 0, 1 or 2, we translate all the 2s to 1s. We select two disjoint subsets of the data (subsetting on SNPs not observations), one with $p = 2000$ to study the $p > n$ case, and the other with $p = 400$ to look at the $p < n$ case.

We study DET by simulating binary responses on the data, with various relationships to the binary predictors. Throughout this section we select the significance level for stage

one of DET via a Bonferroni correction to achieve a family-wise error rate of 0.05. We record two kinds of finds from the DET method — a 'primary find' (pfind) means that a true causal predictor is identified by the first stage of the method, whilst a 'secondary find' (sfind) means that a true causal predictor is contained in the set $\mathcal{J}$ associated with a significant direct effect, and has a probability of at least 0.1 of being a direct effect. A 'false find' (ffind) occurs when a significant direct effect is found but there are no associated primary or secondary finds.

We compare DET with a standard logistic regression with a lasso penalty, where the strength of the penalty is chosen via BIC. We define a 'find' under the standard lasso occurring when a true causal predictor is assigned a non-zero coefficient. Significance testing is not appropriate because the relatively small sample size coupled with the multicollinearity of the dataset means that we do not find coefficients that are significantly different from zero. A lasso 'false find' (ffind) occurs when a non-zero coefficient is assigned to a non-causal predictor. We additionally compare with the 'screen and clean' (S&C) method of Wasserman and Roeder [2009], where the strength of the penalty in the 'screen' stage is chosen via BIC. In Wasserman and Roeder [2009] cross validation is used to determine the penalty for the 'screen' stage — this leads to more variables being carried forward to the 'clean' stage, compared with BIC, and hence more true and false finds. Due to the high multicollinearity in this particular dataset, the increase in false finds was particularly damaging for both the lasso and the 'screen and clean' methods, so using BIC seemed to give more favourable results for these methods. The significance level for the 'screen' stage of the screen and clean procedure is again chosen to achieve a family-wise error rate of 0.05.

It must be noted that, for the lasso and screen and clean methods, a find is usually declared to have occurred when a non-zero co-efficient is found on a predictor highly correlated with the causal predictor. However we are considering the case when it is of interest to recover the causal predictor exactly.

For each of the $p > n$ and $p < n$ cases, we carry out 100 independent simulations, where in each case, causal predictor(s) are randomly selected, and a response is simulated via various relationships to these causal predictor(s). We study here cases of one and two causal predictors, with effect sizes of 10% and 20%. Table 2 gives the results for the $p > n$ case and Table 3 gives the results for the $p < n$ case. The number of finds made by lasso and DET are very similar, despite DET implementing a stringent significance test and lasso merely reporting non-zero coefficients. In addition, the lasso makes a larger number of false finds in general. The screen and clean method achieves similar false find control to DET, but this is at the expense of a far smaller number of true finds.

4 Heart Disease Data

We now illustrate the method on a real dataset. The Coronary Risk-Factor Study [Rousseauw et al., 1983] was carried out in three rural areas in South Africa, in the White Cape region, where incidence of heart disease is particularly high. A subset of the study is analysed extensively in Hastie et al. [2001]. In this subset a binary response is measured, whether or not the subject has heart disease, and 160 cases and 302 controls are collected. Each subject has nine measurements taken as predictors. These are 'sbp' (systolic blood pressure); 'tobacco' (cumulative tobacco); 'ldl' (low density lipoprotein cholesterol); 'adiposity';

Table 2: Comparison of lasso and DET finds for $p > n$ case for various effect sizes, for one and two causal predictors

Effect	0.2	0.1	(0.2,0.2)	(0.1,0.1)
Lasso finds	36	1	95	5
Lasso ffinds	24	5	45	11
S&C finds	17	0	28	0
S&C ffinds	10	7	12	2
DET pfinds	37	2	80	6
DET sfinds	7	1	17	0
DET finds	44	3	97	6
DET ffinds	11	4	19	5

'famhist' (family history of heart disease); 'typea' (type-A behaviour); 'obesity'; 'alcohol' (current alcohol consumption); and 'age' (age at onset, or age of testing for controls). To illustrate DET, we have dichotomized the predictors where necessary, by setting a single threshold level, at an appropriate point where possible: for example, the 'obesity' predictor measures Body Mass Index (BMI) so we we have used 30 as the cut-off point, since persons with a BMI exceeding 30 are classed as obese.

We then carry out five analyses on the dichotomized data: the standard single predictor association test, a standard logistic regression, a logistic regression with lasso penalty, the screen and clean method and the direct effect testing method. Results of the single predictor test, the logistic regression and the screen and clean method are given in Table 4. For the

Table 3: Comparison of lasso and DET finds for $p < n$ case for various effect sizes, for one and two causal predictors

Effect	0.2	0.1	(0.2,0.2)	(0.1,0.1)
Lasso finds	53	6	115	11
Lasso ffinds	27	3	51	9
S&C finds	25	2	48	5
S&C ffinds	9	3	13	4
DET pfinds	49	3	82	13
DET sfinds	7	0	15	1
DET finds	56	3	97	14
DET ffinds	10	4	17	4

screen and clean method, some variables are 'dropped' at the screen stage, so they do not have associated p-values. For the lasso method, four non-zero coefficients were identified — on 'tobacco', 'ldl', 'famhist' and 'age'. For the direct effect testing method, four direct effects were found at the Bonferroni significance level of 0.0056, and the details are in Table 5. The probabilities in Table 5 do not always sum to one, due to rounding and exclusion of predictors with low ($< 0{\cdot}01$) probabilities, using the cut-off rule specified in Section 2.4.

To summarize the findings of the DET analysis, we are virtually certain that 'age', 'famhist' and 'tobacco' have a direct effect on heart disease, this is reflected in the small p-values in both the logistic regression and the single predictor analysis. There is a possible fourth direct effect, and 'tobacco' re-appears as a possible predictor to possess this direct

Table 4: Comparing p-values calculated via the standard single predictor test and a logistic regression, for the heart disease data

Covariate	Single Predictor	Logistic Regression	S& C
age	$1{\cdot}1 \times 10^{-11}$	$9{\cdot}0 \times 10^{-4}$	$3{\cdot}7 \times 10^{-3}$
famhist	$4{\cdot}8 \times 10^{-9}$	$1{\cdot}1 \times 10^{-5}$	$7{\cdot}0 \times 10^{-3}$
ldl	$4{\cdot}4 \times 10^{-7}$	$6{\cdot}9 \times 10^{-2}$	$3{\cdot}4 \times 10^{-2}$
adiposity	$4{\cdot}4 \times 10^{-6}$	$2{\cdot}4 \times 10^{-1}$	dropped
tobacco	$3{\cdot}2 \times 10^{-7}$	$1{\cdot}0 \times 10^{-1}$	$8{\cdot}3 \times 10^{-2}$
typea	$2{\cdot}3 \times 10^{-1}$	$4{\cdot}3 \times 10^{-2}$	dropped
sbp	$8{\cdot}1 \times 10^{-4}$	$2{\cdot}8 \times 10^{-1}$	dropped
alcohol	$1{\cdot}3 \times 10^{-1}$	$7{\cdot}0 \times 10^{-1}$	dropped
obesity	$1{\cdot}3 \times 10^{-1}$	$3{\cdot}3 \times 10^{-1}$	dropped

Table 5: Details from direct effect testing method for heart disease data

Direct Effect p-value	Location	Probability
$4{\cdot}5 \times 10^{-8}$	age	1
$3{\cdot}0 \times 10^{-6}$	famhist	1
$2{\cdot}1 \times 10^{-4}$	tobacco	1
$1{\cdot}3 \times 10^{-3}$	tobacco	0·64
	ldl	0·31
	age	0·02
	typea	0·02
	adiposity	0·01

effect. We interpret this as either evidence that the direct effect is elsewhere so that 'ldl' becomes the most likely origin for the fourth direct effect; an interaction effect; or evidence that this fourth direct effect is in fact a false positive.

5 Discussion

In this paper we have introduced, for binary predictors and response, a method that separates the testing for the presence of a direct effect and the selection of the predictor that produces the effect. This allows, in the first stage, direct effect hypothesis tests to be carried out in the presence of highly correlated predictors without suffering multicollinearity issues. The uncertainty in the assignment of a direct effect to a predictor, caused by the multicollinearity, is taken into account in the second stage, so that the method gives a set of predictors that could represent each direct effect, with probabilities on each predictor in the set. We demonstrate that the method works effectively to find single and multiple direct effects, and compares very favourably with the lasso. Whilst similar methods are available [Meinshausen, 2008], DET is unique in offering a probabilistic assessment of which predictors could be associated with the detected effect.

The second stage of the method can be viewed from a Bayesian perspective, by relaxing assumption 2 given in Section 2.4, and instead placing a discrete prior on pr(X_j true DE). The enforcement of assumption 2 corresponds to a uniform prior.

The method easily handles missing data, provided we use the missing completely at random assumption [Rubin, 1976]. Since we deal with cell counts only, values, i.e. a specific x_{ij}, that are missing at any point can be excluded from the count, and therefore no

imputation is required. The column totals in Table 1 would then depend on j so we would replace s by s_j, and so forth.

One of the shortcomings of the method is that it does not allow for multiple levels of the predictor variables. One way to address this issue is by introducing multiple binary predictors for a single discrete predictor. For example, consider a three level predictor X_j, taking values 0,1 or 2. Then we introduce two binary predictors, X_{j_1} and X_{j_2}. Code $X_{j_1} = 1$ if $X_j \geq 1$ and $X_{j_1} = 0$ if $X_j = 0$; and code $X_{j_2} = 1$ if $X_j = 2$ and $X_{j_2} = 0$ if $X_j \leq 1$. A more general extension that makes use of the multivariate hypergeometric distribution will be investigated in the future.

Another interesting point for future investigation are the connections of the introduced method to Genomic Control [Devlin and Roeder, 1999, Devlin et al., 2001] and Delta Centralisation [Gorroochurn et al., 2006], which are methods used to account for subpopulation structure or other unobserved confounding effects in a dataset, particularly applied in genetic contexts. This is achieved by assuming the better known noncentral χ^2 null distribution in tests of association, with a noncentrality parameter ν that is common to all tests. This begs the question of whether the direct effect testing method can be used in a similar context, and whether additional power is gained by allowing for a different noncentrality parameter for each test.

The most important generalization required for this method, perhaps, is to allow for continuous predictors and response. Whether this is possible remains an open question — we envisage that the main difficulties would be calculation of the matrix E, and whether it is possible to perform parametric hypothesis testing in this case.

References

Y. Aulchenko and M. Struchalin. *GenABEL: genome-wide SNP association analysis*, 2008. R package version 1.4-0.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37: 373–384, 1995.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Elec. Journ. Stat.*, 1:169–194, 2007.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.*, 35(6):2313–2351, 2007.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55 no. 4: 997–1004, 1999.

B. Devlin, K. Roeder, and L. Wasserman. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, 60:155–166, 2001.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32 (2):407–499, 2004.

J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *J. Roy. Stat. Soc.*, 70:849–911, 2008.

P. Gorroochurn, G. A. Heiman, S. E. Hodge, and D. A. Greenberg. Centralizing the non-

central chi-square: a new method to correct for population stratification in case-control association studies. *Genetic Epidemiology*, 30:277–289, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

A. Hoerl and R. Kennard. Ridge regression. In *Encyclopedia of Statistical Sciences*, pages 129–136. New York: Wiley, 1988.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Stat.*, 28:1356–1378, 2000.

B. Levin. Simple improvements on Cornfield's approximation to the mean of a noncentral hypergeometric random variable. *Biometrika*, 71:630–632, 1984.

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

W. Massey. Principal components regression with exploratory statistical research. *J. Am. Statist. Ass.*, 60:234–246, 1965.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. Technical report, Seminar für Statistik, ETH Zürich, 2008.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. Technical report, Seminar für Statistik, ETH Zürich, 2008.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. Technical Report 98/1, Dept. Statistics, Univ. Adelaide, 1998.

J. Pearl. *Causality*. Cambridge University Press, second edition, 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Rousseauw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58(1): 267–288, 1996.

L. Wasserman and K. Roeder. High dimensional variable selection. *Ann. Stat.*, 2009. In press.

H. Wold. *Soft modeling by latent variables: the nonlinear iterative partial least squares approach*. New York: Academic Press, 1975.

H. Zou. The adaptive lasso and its properties. *J. Amer. Stat. Assoc.*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67(2):301–320, 2005.

Appendix

Derivation of Equation 4

First, by definition

$$e_j^k = E(z_j \mid C_k, z_k) = \sigma_j^{-1}\big\{n\mathrm{pr}(\mathrm{X_j} = 0, \mathrm{Y} = 0 \mid \mathrm{C_k}, \mathrm{z_k}) - \mu_\mathrm{j}\big\},$$

where the observation index i is suppressed. Assumption C_k means that X_j is conditionally independent of Y, given X_k, since X_k separates X_{-k} from Y in the graph. Using this conditional independence, we then find

$$\begin{aligned}
\mathrm{pr}(\mathrm{X_j} = 0, \mathrm{Y} = 0 \mid \mathrm{C_k}, \mathrm{z_k}) &= \mathrm{pr}(\mathrm{X_j} = 0, \mathrm{Y} = 0 \mid \mathrm{X_k} = 0, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{X_k} = 0 \mid \mathrm{C_k}, \mathrm{z_k}) \\
&\quad + \mathrm{pr}(\mathrm{X_j} = 0, \mathrm{Y} = 0 \mid \mathrm{X_k} = 1, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{X_k} = 1 \mid \mathrm{C_k}, \mathrm{z_k}) \\
&= \mathrm{pr}(\mathrm{X_j} = 0 \mid \mathrm{X_k} = 0, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{Y} = 0 \mid \mathrm{X_k} = 0, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{X_k} = 0 \mid \mathrm{C_k}, \mathrm{z_k}) \\
&\quad + \mathrm{pr}(\mathrm{X_j} = 0 \mid \mathrm{X_k} = 1, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{Y} = 0 \mid \mathrm{X_k} = 1, \mathrm{C_k}, \mathrm{z_k})\mathrm{pr}(\mathrm{X_k} = 1 \mid \mathrm{C_k}, \mathrm{z_k}),
\end{aligned} \tag{15}$$

We proceed by collecting the first and third terms from each of the above lines, to give

$$\begin{aligned}\mathrm{pr}(\mathrm{X_j}=0,\mathrm{Y}=0\mid \mathrm{C_k},\mathrm{z_k}) &= \mathrm{pr}(\mathrm{X_j}=0,\mathrm{X_k}=0\mid \mathrm{C_k},\mathrm{z_k})\mathrm{pr}(\mathrm{Y}=0\mid \mathrm{X_k}=0,\mathrm{C_k},\mathrm{z_k}) \\ &\quad +\mathrm{pr}(\mathrm{X_j}=0,\mathrm{X_k}=1\mid \mathrm{C_k},\mathrm{z_k})\mathrm{pr}(\mathrm{Y}=0\mid \mathrm{X_k}=1,\mathrm{C_k},\mathrm{z_k}) \\ &= \frac{\gamma_{0,0}a_k}{a_k+b_k}+\frac{\gamma_{0,1}c_k}{c_k+d_k},\end{aligned}$$

where the last line follows from deriving (a_k, b_k, c_k, d_k) from z_k, which we have conditioned on throughout.

Derivation of Equation (8)

Write $p_{j|k} = \mathrm{pr}(X_j \text{ true}|X_k \text{ dec.})$, abbreviating in the obvious way. Now by Bayes' Theorem,

$$p_{j|k} \propto \mathrm{pr}(X_k \text{ dec.}|X_j \text{ true})\mathrm{pr}(X_j \text{ true}).$$

But

$$\begin{aligned}\mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}) &= \mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}, X_k \text{ or } X_j \text{ dec.})\mathrm{pr}(X_k \text{ or } X_j \text{ dec.}|X_j \text{ true}) \\ &= \mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}, X_k \text{ or } X_j \text{ dec.})\big\{\mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}) + \mathrm{pr}(X_j \text{ dec.}|X_j \text{ true})\big\},\end{aligned}$$

then re-arranging gives

$$\mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}) = \frac{\mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}, X_k \text{ or } X_j \text{ dec.})\mathrm{pr}(X_j \text{ dec.}|X_j \text{ true})}{1-\mathrm{pr}(X_k \text{ dec.}|X_j \text{ true}, X_k \text{ or } X_j \text{dec.})}.$$

So that

$$p_{j|k} \propto \mathrm{odds}(X_k \text{ dec.}|X_j \text{ true}, X_k \text{ or } X_j \text{ dec.})\mathrm{pr}(X_j \text{ dec.}|X_j \text{ true})\mathrm{pr}(X_j \text{ true})$$

as required.

Derivation of Equation (11)

Referring to Table 1, if we were interested in the size of the association between X_k and Y, we would estimate this as

$$\begin{aligned} \Pr(Y = 1 \mid X_k = 0) &= \frac{b_k}{t_{0k}} \\ &= \frac{t_{0k} - a_k}{t_{0k}} \\ &= \frac{t_{0k} - \mu_k - \sigma_k z_k}{t_{0k}}, \end{aligned}$$

where we replace a_k by the noncentrality model of Equation (6). Removing the indirect effect part, $\sum_{j \neq k} m_j e_k^j$ then immediately yields $\hat{\alpha}_k$ in Equation (11).

In order to find the expression for $\hat{\beta}_k$, consider

$$\begin{aligned} \Pr(Y = 1 \mid X_k = 1) &= \frac{d_k}{t_{1k}} \\ &= \frac{r - t_{0k} + a_k}{t_{1k}} \\ &= \frac{r - t_{0k} + \mu_k + \sigma_k z_k}{t_{1k}}. \end{aligned}$$

Removing the indirect effect part and subtracting $\hat{\alpha}_k$ then yields the desired result.

Derivation of Equation (12)

Recall that we assume a true direct effect between X_j and Y. We then find

$$\begin{aligned} P_{EF(j,k)} &= \text{pr}\left\{(\mathrm{X_j}, \mathrm{X_k}, \mathrm{Y}) = (0,1,1) \text{ or } (1,0,0) \mid \mathrm{X_j} \neq \mathrm{X_k}\right\} \\ &= \frac{\text{pr}\left\{(\mathrm{X_j}, \mathrm{X_k}, \mathrm{Y}) = (0,1,1)\right\} + \text{pr}\left\{(\mathrm{X_j}, \mathrm{X_k}, \mathrm{Y}) = (1,0,0)\right\}}{\text{pr}\left\{(\mathrm{X_j}, \mathrm{X_k}) = (0,1)\right\} + \text{pr}\left\{(\mathrm{X_j}, \mathrm{X_k}) = (1,0)]\right\}} \\ &= \frac{\gamma_{(1,0)}\alpha_k + \gamma_{(0,1)}(1 - \alpha_k - \beta_k)}{\gamma_{(1,0)} + \gamma_{(0,1)}}, \end{aligned}$$

where the last line is obtained by writing $\mathrm{pr}\left(\mathrm{X_j}, \mathrm{X_k}, \mathrm{Y}\right) = \mathrm{pr}\left(\mathrm{Y} \mid \mathrm{X_j}, \mathrm{X_k}\right) \mathrm{pr}\left(\mathrm{X_j}, \mathrm{X_k}\right)$, and using Equation (10) for the conditional probabilities of Y.